DOCUMENT RESUME

ED 144 403 FL 008 909

AUTHOR Mullen, Karen A.
TITLE Rater Reliability and Oral Proficiency Evaluations. Occasional Papers on Linguistics, No. 1.
INSTITUTION Southern Illinois Univ., Carbondale. Dept. of Linguistics.
PUB DATE Apr 77
NOTE 16p.; Proceedings of the International Conference on Frontiers in Language Proficiency and Dominance Testing (1st, Southern Illinois University, Carbondale, Illinois, April 21-23, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *English (Second Language); Evaluation Methods; Foreign Students; Grammar; Higher Education; Interviews; Language Fluency; *Language Proficiency; *Language Tests; Listening Comprehension; Pronunciation; *Rating Scales; Second Language Learning; *Speech Skills; Student Evaluation; Test Reliability; Vocabulary
IDENTIFIERS *Rater Reliability

ABSTRACT

A study was conducted that was designed to determine if experienced ESL (English as a second language) teachers, working in pairs, can reach the same judgments regarding the oral proficiency of non-native speakers of English and to determine the reliability of such judgments. In addition, the question of whether different sets of judges rate differently is posed. Finally, the study was designed to determine the relative weight given to each component category in predicting the overall proficiency score. Five judges, all graduate students in linguistics who had taught ESL for at least one year, were randomly paired together to form six groups. The 98 subjects were foreign students referred to the University of Iowa for evaluation. The judges were required to rate speakers in five categories of speaking proficiency: (1) pronunciation, (2) grammar; (3) vocabulary; (4) fluency; and (5) auditory comprehension. The results showed that although there were differences in the ratings among the individual judges, an average of the two judges' ratings served as a good estimate of the true rating. The ratings on the listening comprehension, pronunciation, fluency, and overall proficiency scales did not differ from group to group. A significant difference was found, however, for the grammar scale. The overall scale appears to be a composite of the other four scales and, therefore, is the best scale of measurement for oral proficiency. (CFM)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

James E. Redden, Editor

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED144403

Rater Reliability and Oral Proficiency Evaluations

Karen A. Mullen
University of Iowa

Testing speaking proficiency has been of special interest in the field of second language learning and it has generally been recognized that the best way to test for oral proficiency is to have a subject speak. However, the issue of what to test and how to test for it has remained. Five basic components of speaking skill have been proposed for analyses of the speech process by Harris (1969): pronunciation, grammar, vocabulary, fluency, and auditory comprehension. Of these five, fluency has been said to be the easiest to assess since it focuses only on pauses, backtracking, and fillers in the speech flow. On the other hand, pronunciation has been said to be difficult to judge since criteria for judgment may vary from one person to another; some listeners may be able to readily decode foreign accents; others may be judging comprehensibility rather than phonemic or allophonic accuracy. Moreover, three of these components (grammar, vocabulary, and auditory comprehension) can be tested in ways other than asking a subject to speak. However, evidence of a formal knowledge of grammar and vocabulary does not guarantee that it will be applied in actual speech production.

The question of how to test for speaking proficiency has usually been decided in favor of the interview. It has certain advantages: it can be conducted rather quickly, it most resembles real-life speaking situations, and it can be adjusted up or down depending upon the speaker's demonstrated proficiency. It, however, has been open to the criticism that the measures derived from such a test tend to have rather low reliability. Some have suggested that a tolerable degree of reliability can be achieved if behavioral statements are used as a standard for each scale judgment, if judges are trained for their task, and if at least two judgments are pooled for each interview. The purpose of this paper is to report a study designed to determine if experienced ESL teachers, working in pairs, can reach the same judgments regarding the oral proficiency of non-native speakers of English and to determine the reliability of such judgments. In addition, the question of whether differents sets of judges rate differently is posed. Finally, the study was designed to determine the relative weight given to each component category in predicting the overall proficiency score. Specifically, the hypotheses are as follows:

1. The ratings assigned by judge 1 in a specific category of oral proficiency are not significantly different from those of judge 2, when judge 1 and judge 2 are rating the same subject.
2. The ratings assigned by one pair of judges to a group of speakers are not significantly different from those assigned by other pairs of judges to other groups of speakers.
3. Each category is given weight by the judges in their assessment of a subject's overall proficiency.

FL008909

Experimental Design

To test hypothesis 1, a single-factor experimental design having repeated measures was chosen. The F-statistic based upon the mean square between judges divided by the mean square of the judge-subject interaction was computed to test the hypothesis of no significant differences between judges. Reliability coefficients (unbiased) were calculated based upon the number of subjects in the sample, the number of judges (2), the mean square between subjects, and the mean square within subjects (Winer, 287). To test hypothesis 2, a two-factor experimental design having repeated measures on one factor was chosen. The F-statistic based upon the mean square between groups divided by the mean square of subjects within groups was computed to test the hypothesis of no significant differences among groups. To test hypothesis 3, the regression coefficients of a multiple regression equation were calculated and a F-test for a coefficient of zero was performed.

Procedure

Judges were required to rate speakers in five categories of speaking proficiency: listening comprehension, pronunciation, fluency, control over English structure, and overall speaking proficiency. These scales were labelled vertically on a rating form. Beside each of the five scales was placed a double horizontal line equally divided into five contiguous compartments labeled from left to right: poor, fair, good, above average, excellent. The judges were instructed to put an X in the box best characterizing the speaker's proficiency with regard to each to the five scales or to put an X on the line between boxes if the subject's proficiency seemed to be between the two labelled areas. A set of guidelines for deciding what level of proficiency to assign was explained to the judges before they interviewed and it was at hand for reference after each interview. Each pair of judges was instructed to ask the subject a few questions in order to get the interview started -- what the subject's name was, where he was from, how long he had been in the U.S., where he had studied, what his major field of interest was, etc. These were intended to put the subject at ease as quickly as possible. The interviewers were instructed to begin by speaking at a normal rate and to slow down later if it became apparent that the subject could not follow. In lieu of grossly distorting their speech, the interviewers rephrased their questions by simplifying the sentence structure and vocabulary. If it became evident that the subject was able to do well in a rather informal type of interview, the judges were prepared to shift the questioning to a level more like that which subjects would encounter in an informal academic setting. They would then ask such questions as: what interests you most about your field of study, what are some important questions yet to be answered, what special subfields exist, what are the relationships between them, etc. The intent was to simulate a conversation and the next question in the sequence, of course, depended on what had preceded. One of the aims was to vary the questioning from subject to subject so that later subjects could not be apprised of the questions beforehand. Raters were instructed to contribute to the conversation occasionally but to control their speaking in such a way that it could be used to assess the subject's listening comprehension. They were encouraged to give the subject every opportunity to demonstrate his speaking proficiency. After each interview, lasting approximately fifteen minutes, each judge evaluated the subject

without consulting the other. Evaluations made by the individual judges were later converted to a numerical value of 1=poor, 2=between poor and fair, 3=fair, 4=between fair and good, 5=good, 6=between good and above average, 7=above average, 8=between above average and excellent, and 9=excellent.

Judges

Five judges participated in this study. They were randomly paired together to form six groups. All judges were graduate students in linguistics. They had completed courses in phonetics, syntactic and phonological analysis, and TESL methodology. They all had taught ESL for at least one year. All had been instructed on how to use the rating form and guidelines, and all had participated in such interviewing before. None of the judges had previously met the subjects they interviewed.

Subjects

The ninety-eight subjects in this study had been referred to the University of Iowa Department of Linguistics for a proficiency evaluation by either the foreign admissions officer, the foreign student advisors, or the student's academic advisor. Most of the subjects were new to the university and were referred because their TOEFL socres were below 550. Some were evaluated because the foreign student advisors had noted a lack of facility in spoken English although their TOEFL scores were not below 550. The purpose of the evaluation was to determine whether additional work in English and a reduced academic program might be recommended for the student.

Results

Table 1 shows the results of a two-factor analysis of variance having repeated measures on one factor. A separate analysis of variance is reported for each of the five scales of speaking proficiency. The F-statistic for differences between groups is significant at the .05 level for listening, pronunciation, and overall speaking proficiency. It is significant at the .01 level for grammar. It is not significant at either of these two levels for fluency. The F-statistic for differences between judges is significant at the .01 level for all scales.

Table 2 shows the results of a single-factor analysis of variance having repeated measures for each of the six groups and for each of the five rating scales. The F-statistic for a difference between judges' listening ratings is significant at the .05 level for three pairs of judges (groups 4,5, and 6), between pronunciation ratings for two pairs (groups 1 and 2), between fluency ratings for one pair (group 2), between grammar ratings for three pairs (groups 1, 2, and 3), and between overall ratings for two pairs (groups 2 and 3). For every pair of judges, there was a significant difference in ratings on at least one scale. Additionally, one pair of judges (group 2) showed a significant difference in ratings on all but one scale. Two pairs (groups 5 and 6) showed a significant difference on just one scale, that of listening comprehension. At the .01 level, there is no significant difference between judges across rating scales for three sets of judges (groups 3, 5, and 6). Group 2 shows a significant difference between judges on every scale except listening.

		Listening			Pronunciation			Fluency		
Source of Variation	df	SS	MS	F	SS	MS	F	SS	MS	F
Between groups	5	58.0043	11.6008	2.45*	38.5367	7.7073	2.78*	23.8785	4.7757	1.13
Between subjects within groups (pooled)	92	435.9752	4.7388		255.4071	2.7761		388.0040	4.2174	
Between judges within groups (pooled)	6	18.5267	3.0877	4.85**	9.3108	1.5518	3.38**	18.7402	3.1233	5.55**
Subjects x judges (pooled)	92	58.4732	.6355		42.1891	.4585		51.7596	.5626	

		Grammar			Overall		
Source of Variation	df	SS	MS	F	SS	MS	F
Between groups	5	51.5835	10.3167	4.48**	46.6350	9.3270	2.95*
Between subjects within groups (pooled)	92	212.0899	2.3053		291.0384	3.1634	
Between judges within groups (pooled)	6	7.2596	1.2099	30.32**	7.6784	1.2979	4.15**
Subjects x judges pooled)	92	36.7403	.0399		28.3215	.3078	

* significant at $p < .05$
** significant at $p < .01$

Table 1. An Analysis of Variance of Performance Scores of Six Groups of Subjects Tested under Different Pairs of Judges on Five Scales of Speaking Proficiency.

			Listening		Pronunciation		Fluency		Grammar		Overall	
Group	Source of Variation	df	MS	F	MS	F	MS	F	MS	F	MS	F
1	Between subjects	14	6.561	15.31**	3.557	5.15**	4.461	4.85*	1.633	7.98**	3.657	10.97**
	Between judges	1	.000	.00	3.333	4.83*	1.633	1.78	2.133	10.42**	.833	2.50
	Judges x subjects	14	.428		.690		.919		.204		.333	
2	Between subjects	16	2.908	6.88**	3.132	8.69**	3.132	8.69**	1.242	3.93**	1.889	10.28**
	Between judges	1	.735	1.74	4.235	11.76**	14.235	39.51**	2.941	9.30**	3.558	19.36**
	Judges x subjects	16	.422		.360		.360		.316		.183	
3	Between subjects	9	3.311	2.26	1.888	5.31*	2.338	2.23	1.911	7.82**	2.022	8.27**
	Between judges	1	1.800	1.23	.800	2.25	.050	.05	1.800	7.36*	1.800	7.36*
	Judges x subjects	9	1.466		.355		1.050		.244		.244	
4	Between subjects	16	6.341	11.27**	3.867	13.84**	5.746	15.48**	2.680	3.18*	4.753	16.79**
	Between judges	1	8.500	15.11**	.029	.11	1.058	2.85	.029	.03	.470	1.66
	Judges x subjects	16	.562		.279		.371		.841		.283	
5	Between subjects	24	4.578	7.29**	1.916	3.18*	4.958	10.02**	2.795	7.73**	2.911	6.64**
	Between judges	1	3.920	6.24*	.020	.03	1.620	3.27	.320	.88	.980	2.24
	Judges x subjects	24	.628		.603		.495		.361		.438	
6	Between subjects	13	4.340	6.69**	2.354	6.64**	3.340	7.41**	3.244	9.45**	3.497	13.12**
	Between judges	1	3.571	5.51*	.892	2.52	.142	.32	.035	.10	.035	.13
	Judges x subjects	13	.648		.354		.450		.343		.266	

* significant at $p < .05$

** significant at $p < .01$

Table 2. Analyses of Variance of Performance Scores (one for each of six groups of tested under a different pair of judges) on Five Scales of Speaking Proficiency.

Table 3 shows the differences in judges' mean ratings for each group. It is visually apparent from the graphs that some lines rise more sharply than others and that the distance between the lowest and the highest line is great enough to result in a significant difference in groups at the .05 level. Group 2 stands out as the most deviant with regard to differences between judges and with regard to mean ratings when compared to other groups.

Table 4 shows the reliability coefficients (unbiased) for each pair of judges for all scales of oral proficiency. The coefficients range from a low of .43 to a high of .93. Table 4 also shows the relationship between the F-statistic and the reliability coefficients. Some cases show a significant difference in judges' ratings and a high reliability coefficient (for example, listening ratings in group 4). The reverse is also apparent (for example, fluency ratings in group 3). In addition, some cases show no significant difference in judges' ratings and a high reliability coefficient (for example, listening ratings in group 1) as well as the reverse (for example, grammar ratings in group 2). Since the reliability coefficient is a measure of the degree to which the average error of measurement is zero and indicates how good an average of the two judges' ratings is as an estimate of the subject's true rating, in cases where there is a significant difference in judges' ratings and a high reliability (above .70), one may conclude that an average of the two ratings is a better estimate of the subject's true rating than one alone. In cases where there is no significant difference in judges' ratings and a high reliability, one may conclude that either of the two ratings is as good an estimate of the subject's true rating as an average of the two is. Where there is a significant difference in judges' ratings and a low reliability, we may conclude that an average of the two judges' ratings is not a good estimator of the subject's true rating. Likewise, where there is no significant difference in judges' ratings and a low reliability, it is evident that the average error of measurement is sufficient to reduce the extent to which an average of the two ratings is a good estimator of the subject's true rating. The overall scale appears to provide the most uniform reliability coefficients across groups of subjects. Since a significant difference in judges' ratings is evident, it is best to use an average of the two judges' ratings as the best estimate.

As shown in Table 5, the correlation between scale ratings is high. However, a stepwise multiple regression analysis reveals that as the variables are added one by one to the equation for predicting the overall proficiency rating, each addition to the equation significantly reduces the amount of unpredicted variance. Table 6 shows these results. The beta coefficients in the multiple regression equation to predict the overall rating from the four other scales indicates that if the listening rating were increased by one unit and the other ratings remained constant, the expected change in the overall proficiency rating would be .23. If the grammar rating were increased by one (and the other ratings remained constant), the change in the overall proficiency rating would be .29. Parallel conclusions may be reached for the influence of an increase in the pronunciation or fluency ratings on the overall proficiency rating. If all four ratings were increased by one unit, the change in the overall proficiency rating would be very close to one unit as well. This indicates that each of the scales is weighted

		Listening				Pronunciation				Fluency				Grammar				Overall			
		Judge 1		Judge 2		Judge 1		Judge 2		Judge 1		Judge 2		Judge 1		Judge 2		Judge 1		Judge 2	
group	n	mean	S.D.	mean	S.D.	mean	S.D.	mean	S.D.	mean	S.D.	mean	S.D.	mean	S.D.	mean	S.D.	mean	S.D.	mean	S.D.
1	15	6.06	1.75	6.06	1.98	5.46	1.55	6.13	1.35	5.73	1.79	6.20	1.47	5.46	.99	6.00	.92	5.73	1.48	6.06	1.33
2	17	7.47	1.46	7.76	1.09	6.41	1.37	7.11	1.26	6.11	1.49	7.41	1.12	6.76	.90	7.35	.86	6.76	1.09	7.41	.93
3	10	6.10	2.02	6.70	.82	5.30	1.25	5.70	.82	5.60	1.17	5.70	1.41	5.50	1.26	6.10	.73	5.40	1.26	6.00	.81
4	17	5.52	1.69	6.52	2.00	5.41	1.37	5.47	1.50	5.64	1.65	6.00	1.83	5.52	1.50	5.58	1.12	5.58	1.66	5.82	1.50
5	25	6.04	1.45	6.60	1.75	5.68	1.21	5.72	1.02	5.72	1.51	6.08	1.77	5.64	1.31	5.80	1.19	5.68	1.34	5.96	1.24
6	14	6.28	1.58	7.00	1.56	5.50	1.22	5.85	1.09	6.07	1.32	6.21	1.42	5.85	1.16	5.92	1.49	5.92	1.43	6.00	1.30

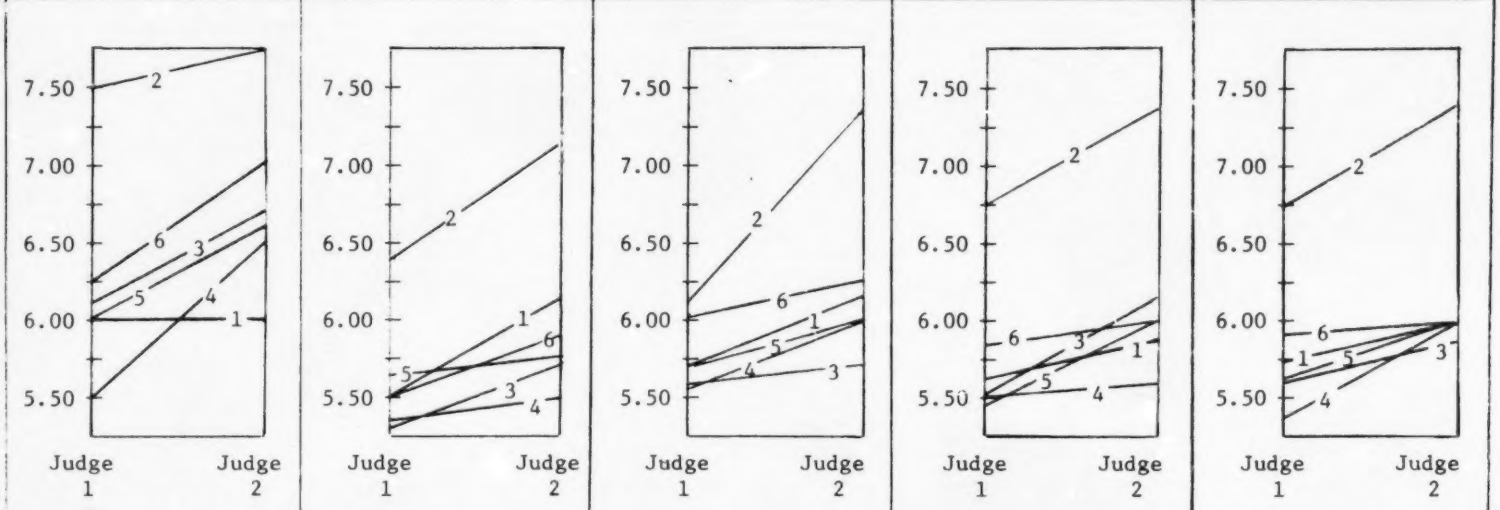

Table 3. Means, Standard Deviations, and Plots of Means of Performance of Six Groups of Subjects Tested under Different Pairs of Judges on Five Scales of Speaking Proficiency.

		Listening		Pronunciation		Fluency		Grammar		Overall	
Group	n	F	r_2	F	r_2	F	r_2	F	r_2	F	r_2
1	15	.00	.93	4.83	.72	1.78	.75	10.42**	.76	2.50	.88
2	17	1.74	.82	11.76**	.79	39.51**	.57	9.30**	.57	19.36**	.77
3	10	1.23	.43	2.25	.74	.05	.50	7.36*	.74	7.36*	.75
4	17	15.11**	.82	.11	.92	2.85	.92	.03	.66	1.66	.93
5	25	6.24*	.82	.03	.67	3.27	.88	.88	.86	2.24	.83
6	14	5.51*	.77	2.52	.81	.32	.85	.10	.88	.13	.92

* significant at $p < .05$
** significant at $p < .01$

Table 4. A Comparison of the F-statistics from Analyses of Variance (from table 2) and Reliability Coefficients of Judge Pairs.

140

Scale	2	3	4	5
1 Listening	.79	.85	.79	.90
2 Pronunciation		.77	.78	.88
3 Fluency			.80	.89
4 Grammar				.89
5 Overall Proficiency				

Table 5. Correlations of Ratings on the Five Scales of Speaking Proficiency.

Source	df	SS	MS	F	r^2	Beta	S.E.(B)	F(for B)
Regression	1	300.060	300.060	790.78**	.803			
Listening	1	300.060	300.060	790.78**		.90	.025	790.78**
Residual	194	73.612	.379					
Regression	2	332.739	166.369	784.40**	.890			
Listening	1	300.060				.51	.031	168.16**
Grammar	1	32.674	32.674	154.07**		.49	.043	154.07**
Residual	193	40.934	.212					
Regression	3	343.842	114.610	737.40**	.920			
Listening	1	300.060				.36	.030	91.99**
Grammar	1	32.674				.36	.040	96.38**
Pronunciation	1	11.091	11.091	71.37**		.31	.038	71.37**
Residual	192	29.841	.155					
Regression	4	348.994	87.248	675.25**	.933			
Listening	1	300.060				.23	.032	34.52**
Grammar	1	32.674				.29	.038	67.81**
Pronunciation	1	11.091				.28	.035	67.60**
Fluency	1	5.162	5.162	39.95**		.24	.034	39.95**
Residual	191	24.678	.129					

** significant at $p < .01$

Table 6. A Stepwise Regression Analysis with the Overall Proficiency Scale as the Dependent Variable.

approximately equally in the determination of the overall proficiency rating. The F-statistic for B in Table 6 indicates that the hypothesis of no weight given to each of the variables is to be rejected.

Discussion

Two of the hypotheses outlined at the beginning of the study are rejected by the results of the analysis of variance on performance ratings by pairs of judges in each of the six groups. Ratings assigned by one judge are significantly different from ratings assigned by the other in six out of thirty cells at the .01 level and eleven out of thirty cells at the .05 level. However, the reliability coefficients for four out of the six significant-difference cases at the .01 level are above .70 and the reliability coefficients for nine out of the eleven significant-difference cases at the .05 level are above .70. Therefore, an average of the two judges' ratings will serve as a good estimator of the true rating. This means that it is essential that there be more than one judge per subject.

There is a significant difference at the .05 level in the ratings assigned by one pair of judges to a group of speakers when compared to the ratings assigned by another pair of judges to another group of speakers. The reasons for this may be due to a difference in judgments made by the raters or a difference in the groups themselves. Since the design of the experiment did not control of homogeneity of groups, it is impossible to determine the source of this difference. However, at the .01 level, there is no significant difference in assigned ratings among the groups on four out of the five scales. We may conclude that the ratings assigned by the judges on the listening, pronunciation, fluency, and overall proficiency scales do not differ from group to group and that with regard to these scales and these raters, the six groups were fairly homogeneous. For the grammar scale, the significant difference in groups may be due to a real difference in performance in grammar among the groups or to a difference in the application of the scale from one set of judges to another.

Each of the four subscales is given weight in the determination of the overall proficiency rating. In addition, using all four scales to predict the overall rating yields significantly less unexplained variance in the overall scale than does using the three best scales, or the two best scales, or the best single scale. Moreover, the scales are given approximately equal weight in the prediction equation and the relationship between the four scales and the overall scale is very nearly perfectly linear. Since the overall scale shows high reliability coefficients across all groups and since it appears to be a composite of the other four scales, it is the best scale of measurement for oral proficiency.